Journal of Instrumentation

Particle identification using Boosted Decision Trees in the Semi-Digital Hadronic Calorimeter prototype

Jinst

Published by IOP Publishing for Sissa Medialab

Received: *April 8, 2020*
Accepted: *July 19, 2020*
Published: *October 13, 2020*

Particle identification using Boosted Decision Trees in the Semi-Digital Hadronic Calorimeter prototype

2020 JINST 15 P10009

The CALICE collaboration

D. Boumediene,[a] A. Pingault,[b] M. Tytgat,[b] B. Bilki,[c] D. Northacker,[c] Y. Onel,[c] G. Cho,[d] D.-W. Kim,[d] S.C. Lee,[d] W. Park,[d] S. Vallecorsa,[d] Y. Deguchi,[e] K. Kawagoe,[e] Y. Miura,[e] R. Mori,[e] I. Sekiya,[e] T. Suehara,[e] T. Yoshioka,[e] L. Caponetto,[f] C. Combaret,[f] R. Ete,[f,1] G. Garillot,[f] G. Grenier,[f] J.-C. Ianigro,[f] T. Kurca,[f] I. Laktineh,[f] B. Liu,[f,o,2] B. Li,[f] N. Lumb,[f] H. Mathez,[f] L. Mirabito,[f] A. Steen,[f,3] E. Calvo Alamillo,[g] M.C. Fouz,[g] J. Marin,[g] J. Navarrete,[g] J. Puerta Pelayo,[g] A. Verdugo,[g] F. Corriveau,[h] M. Chadeeva,[i] M. Danilov,[i,4] L. Emberger,[j] C. Graf,[j] L.M.S. de Silva,[j] F. Simon,[j] C. Winter,[j] J. Bonis,[k] D. Breton,[k] P. Cornebise,[k] A. Gallas,[k] J. Jeglot,[k] A. Irles,[k] J. Maalmi,[k] R. Pöschl,[k] A. Thiebault,[k] F. Richard,[k] D. Zerwas,[k] M. Anduze,[l] V. Balagura,[l] V. Boudry,[l] J.-C. Brient,[l] E. Edy,[l] F. Gastaldi,[l] R. Guillaumat,[l] F. Magniette,[l] J. Nanni,[l] H. Videau,[l] S. Callier,[m] F. Dulucq,[m] Ch. de la Taille,[m] G. Martin-Chassard,[m] L. Raux,[m] N. Seguin-Moreau,[m] J. Cvach,[n] M. Janata,[n] M. Kovalcuk,[n] J. Kvasnicka,[n] I. Polak,[n] J. Smolik,[n] V. Vrba,[n] J. Zalesak,[n] J. Zuklin,[n] Y.Y. Duan,[o] S. Li,[o] J. Guo,[o] J.F. Hu,[o] F. Lagarde,[o] Q.P. Shen,[o] X. Wang,[o] W.H. Wu,[o] H.J. Yang,[o] Y.F. Zhu,[o] L. Emberger,[j] C. Graf,[j] F. Simon[j] and C. Winter[j]

[a] *Université Clermont Auvergne, Université Blaise Pascal, CNRS/IN2P3, LPC, 4 Av. Blaise Pascal, TSA/CS 60026, F-63178 Aubière, France*

[b] *Ghent University, Department of Physics and Astronomy, Proeftuinstraat 86, B-9000 Gent, Belgium*

[c] *University of Iowa, Dept. of Physics and Astronomy, 203 Van Allen Hall, Iowa City, IA 52242-1479, U.S.A.*

[d] *Gangneung-Wonju National University, Gangneung 25457, South Korea*

[e] *Department of Physics and Research Center for Advanced Particle Physics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan*

[1] Now at DESY.
[2] Corresponding author.
[3] Now at NTU.
[4] Also at MIPT.

https://doi.org/10.1088/1748-0221/15/10/P10009

[f] *Univ. Lyon, Univ CLaude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, F-69622 Villeurbanne, France*
[g] *CIEMAT, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas, Madrid, Spain*
[h] *Department of Physics, McGill University, Ernest Rutherford Physics Bldg., 3600 University Ave., Montréal, Québec H3A 2T8, Canada*
[i] *P.N. Lebedev Physical Institute of the Russian Academy of Sciences, 53 Leninsky prospekt, Moscow 119991, Russia*
[j] *Max-Planck-Institut für Physik, Föhringer Ring 6, D-80805 Munich, Germany*
[k] *Universit Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France*
[l] *Laboratoire Leprince-Ringuet (LLR) — CNRS, École polytechnique, Institut Polytechnique de Paris, F-91128 Palaiseau, France*
[m] *Laboratoire OMEGA — École Polytechnique-CNRS/IN2P3, F-91128 Palaiseau, France*
[n] *Institute of Physics, The Czech Academy of Sciences, Na Slovance 2, CZ-18221 Prague 8, Czech Republic*
[o] *Tsung-Dao Lee Institute, Institute of Nuclear and Particle Physics, School of Physics and Astronomy, Shanghai Jiao Tong University, Key Laboratory for Particle Physics, Astrophysics and Cosmology (Ministry of Education), Shanghai Key Laboratory for Particle Physics and Cosmology, 800 Dongchuan Road, Shanghai, 200240, P.R. China*

E-mail: b.liu@ipnl.in2p3.fr

Abstract: The CALICE Semi-Digital Hadronic CALorimeter (SDHCAL) prototype using Glass Resistive Plate Chambers as a sensitive medium is the first technological prototype of a family of high-granularity calorimeters developed by the CALICE collaboration to equip the experiments of future leptonic colliders. It was exposed to beams of hadrons, electrons and muons several times in the CERN PS and SPS beamlines between 2012 and 2018. We present here a new method of particle identification within the SDHCAL using the Boosted Decision Trees (BDT) method applied to the data collected in 2015. The performance of the method is tested first with Geant4-based simulated events and then on the data collected by the SDHCAL in the energy range between 10 and 80 GeV with 10 GeV energy steps. The BDT method is then used to reject the electrons and muons that contaminate the SPS hadron beams. The rejection power of the new method is estimated to be as high as 99.0% for the muons and 99.4% for the electrons associated to a pion selection efficiency of about 95.0%.

Keywords: Analysis and statistical methods; Calorimeter methods; Performance of High Energy Physics Detectors; Gaseous detectors

ArXiv ePrint: 2004.02972

2020 JINST 15 P10009

Contents

1 Introduction **1**

2 Monte Carlo samples and beam data samples **3**

3 Particle identification using Boosted Decision Trees **4**
- 3.1 BDT input variables 4
- 3.2 The two approaches to build the BDT-based classifier 6
  - 3.2.1 MC training approach 7
  - 3.2.2 Data training approach 9

4 Results **11**

5 Conclusion **14**

2020 JINST 15 P10009

1 Introduction

The Semi-Digital Hadronic CALorimeter (SDHCAL) [1] is the first of a series of technological high-granularity prototypes developed by the CALICE collaboration. These technological prototypes have their readout electronics embedded in the detector and they are power-pulsed to reduce the power consumption in experiments proposed within the International Linear Collider (ILC) project [2]. The mechanical structure of these prototypes is part of their absorber. All these aspects increase the compactness of the calorimeters and improve their suitability to apply Particle Flow Algorithm (PFA) techniques [3–5]. The SDHCAL is comprised by 48 active layers, each of them equipped with a 1 m $\times$ 1 m Glass Resistive Plate Chamber (GRPC) and an Active Sensor Unit (ASU) of the same size hosting on one face (the one in contact with the GRPC) pickup pads of 1 cm $\times$ 1 cm and 144 HARDROC2 ASICs [6] on the other face. The GRPC and the ASU are assembled within a cassette made of two stainless steel plates, 2.5 mm thick each. The 48 cassettes are inserted in a self-supporting mechanical structure made of 51 plates, 15 mm thick each, of the same material as the cassettes, bringing the total absorber thickness to 20 mm per layer. The empty space between two consecutive plates is 13 mm to allow the insertion of one cassette of 11 mm thickness. The HARDROC2 ASIC has 64 channels to read out 64 pickup pads. Each channel has three parallel digital circuits whose parameters can be configured to provide 2-bit encoded information indicating if the charge seen by each pad has passed any of the three different thresholds associated to each digital circuit. This multi-threshold readout is proposed to improve on the energy reconstruction of hadronic showers at high energy ($>$ 30 GeV) with respect to the simple binary readout mode as explained in ref. [7]. A picture of the SDHCAL prototype is shown in figure 1.

The SDHCAL was exposed several times to different kinds of particle beams in the CERN PS and SPS beamlines between 2012 and 2018. The energy reconstruction of hadronic showers within

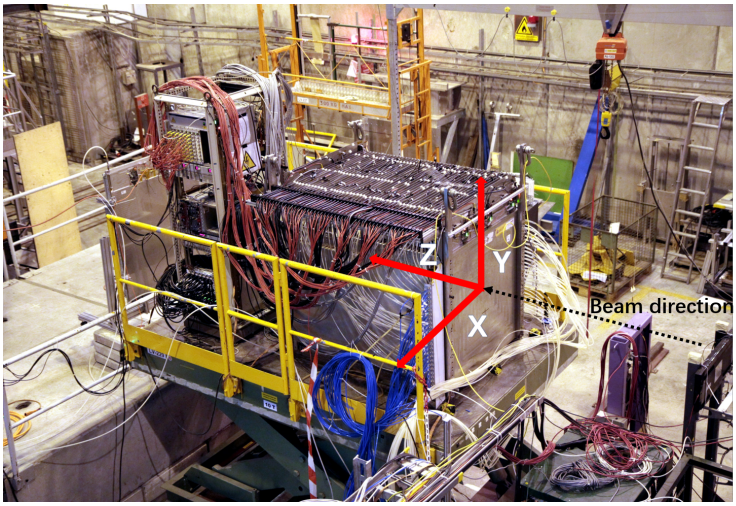

Figure 1. A picture of the SDHCAL prototype on the SPS H2 beamline. The coordinates axes used in the analysis are drawn as well as the beam axis and its direction.

the SDHCAL using the associated number of fired pads with multi-threshold readout information is presented in ref. [7]. The contamination of the SPS hadron beams such as electrons and muons and the absence of Cherenkov counters during data taking require the use of the event's topology to select hadronic events before reconstructing their energy. Although the rejection of muons based on the average number of hits per crossed layer is efficient, the rejection of electrons is more difficult because some hadronic showers (in particular at low energy) behave in similar way as the electromagnetic ones. To reject the electron events, the analysis presented in ref. [7] requires the shower to start after the fifth layer. Almost all of the electrons are expected to start showering before crossing the equivalent of 6 radiation lengths (X_0).[1] Although this selection is found to have no impact on the hadronic energy reconstruction, it keeps only hadrons that shower after about 0.6 interaction length (λ_I) of pions and thus reduces the amount of the hadronic showers available for analysis by about 54%.

In this paper we explore another method to reject electron and muon contaminations that is not based on the shower start requirement and thus provide a larger sample for the energy reconstruction study. The new method is based on Boosted Decision Trees (BDT) [8, 9], a part of so-called Muti-Variate Analysis (TMVA) technique [10]. In the BDT, different variables associated to the topology of the event are exploited in order to distinguish between the hadronic and the electromagnetic showers, and also to identify muons including radiative ones that may exhibit a shower-like shape. In this paper, section 2 introduces the simulation and beam data samples which are used to study the performance of both the BDT and the standard method described in ref. [7]. Section 3 describes the selected input variables of BDT and the two approaches to build the classifier of BDT. Section 4 presents the results of the hadron selection using BDT. Finally, section 5 gives the conclusion.

[1]The longitudinal depth of the SDHCAL prototype layer is about 1.2 X_0.

2020 JINST 15 P10009

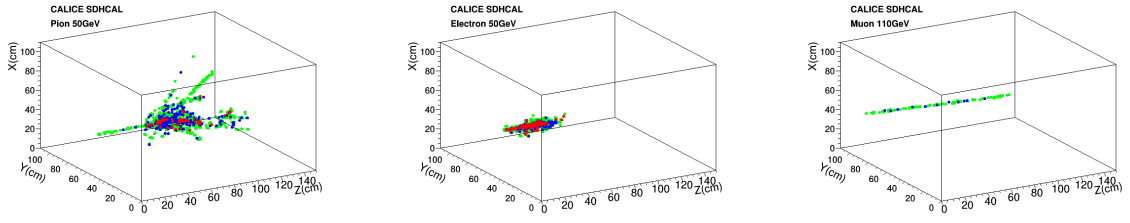

Figure 2. Event displays of a 50 GeV pion in the SDHCAL (left), of a 50 GeV electron (middle) and of a 120 GeV muon (right). Hits that pass the first threshold are depicted in green. Hits that pass the second in blue while the those that pass the third threshold are in red.

2 Monte Carlo samples and beam data samples

The SDHCAL prototype was exposed to pions, muons and electrons in the SPS of CERN in October 2015. In order to avoid GRPC saturation problems at high particle rate, only runs with a particle rate smaller than 1000 particles/spill are selected for the analysis. In these conditions, pion events at several energy points (10, 20, 30, 40, 50, 60, 70, 80 GeV) and muon events of 110 GeV were collected as well as electron events of 10, 20, 30, 40, 50 GeV. While the electron and muon beams are rather pure, the pion beams are contaminated by two sources. One is the electron contamination, despite the use of a lead filter to reduce their contribution. The other is the muon contamination resulting from pions decaying before reaching the prototype. All the active layers of the SDHCAL prototype were operational during this beam test except the layer number 34. Although this layer was physically present, its readout system was switched off due to an electronics problem when the SDHCAL was exposed to the pion beam. Typical pion, electron and muon events are shown in figure 2.

To apply the BDT method, six variables are selected and used in the Toolkit for MultiVariate data Analysis (TMVA) package [10] to build the decision tree.

To study the performance of the BDT method, we use the Geant4.9.6 Toolkit package [11] associated to the FTF-BIC[2] [12, 13] physics list to generate pion, electron and muon events under the same conditions as in the beam test at CERN-SPS beamline. For the training of the BDT, 10k events for each energy point from 10 GeV to 80 GeV with a step of 10 GeV for pions, muons and electrons were produced. In total, 160k events of pions, 160k events of electrons as well as 120k events of muons are used for this study.

The same amount of events of each species is produced and used to test the BDT method at the same time. Finally, the pure (> 99.5%) electron and muon data samples[3] are used as validation sets.

In order to render the particle identification independent of the energy of the different species and thus to extend the method applied here to a larger scope than the beam purification, the pion samples of different energies are mixed before using the BDT technique. The same procedure is applied for muon and electron samples.

[2]The FTF model is based on the Fritiof description of string excitation and fragmentation. The BIC model uses Geant4 binary cascade for primary protons and neutrons with energies below 10 GeV. It describes the production of secondary particles produced in interactions of protons and neutrons with nuclei.

[3]The purity of these samples is provided by the SPS electron and muon beams.

2020 JINST 15 P10009

3 Particle identification using Boosted Decision Trees

Thanks to the high granularity of the SDHCAL, we can use the MVA methods to mine the information of the shape of electromagnetic and hadronic shower to classify muons, electrons and pions. The BDT method is one of the widely used MVA methods to perform such classification tasks. The BDT is a model that combines many less selective decision trees[4] into a strong classifier to achieve a much better performance than single decision tree.

3.1 BDT input variables

The six variables we use to distinguish hadronic showers from electromagnetic showers and from muons are described below. A common right-handed coordinate system is used throughout the SDHCAL whose 48 layers were placed perpendicular to the incoming beams. The origin of the system is defined as the center of the first of the 48 SDHCAL's layers (The x-y plane is parallel to the SDHCAL layers and referred to as the transverse plane while the z-axis runs parallel to the incoming beam as indicated in figure 1.

- **First layer of the shower (Begin):** the probability of a particle to interact in the calorimeter depends on the particle nature and the calorimeter material properties. The distribution of the coordinate z of the layer in which the first inelastic interaction takes place, follows an exponential law. It is proportional to $\exp(-\frac{z}{X_0})$ for electrons and to $\exp(-\frac{z}{\lambda_I})$ for pions, where X_0 and λ_I are the effective radiation length and nuclear interaction length for the SDHCAL material composition, respectively. To define the first layer in which the shower starts we look for the first layer along the incoming particle direction which contains at least 4 fired pads. To eliminate fake shower starts due to accidental noise or a locally high multiplicity, the following 3 layers after the first one are also required to have more than 4 fired pads in each of them. Particles crossing the calorimeter without interaction are assigned the value of 48, which is the case for most of the muons in the studied beam except the radiative ones. Figure 3 shows the distribution of the first layer of the shower in the SDHCAL prototype for pions, electrons and muons as obtained from the simulation and data.

- **Number of tracks segments in the shower (TrackMultiplicity):** applying the Hough Transform (HT) technique to single out the tracks in each event as described in ref. [15], we estimate the number of tracks segments in the pion, electron and muon events. A HT-based segment candidate is considered as a track segment if there are more than 6 aligned hits with not more than one layer separating two consecutive hits. Electron showers feature almost no track segment while most of the hadronic showers have at least one. For muons, except for some radiative muons, only one track is expected as can be seen in figure 4.

- **Ratio of shower layers over total hit layers (NinteractingLayers/NLayers):** this is the ratio between the number of layers in which the Root Mean Square (RMS) of the hits' position in the x-y plane exceeds 5 cm in both x and y directions and the total number of layers with at least one fired pad. It allows, as can be seen in figure 5, an easy discrimination of muons

[4] A decision tree takes a set of input variables and splits input data recursively based on those variables.

2020 JINST 15 P10009

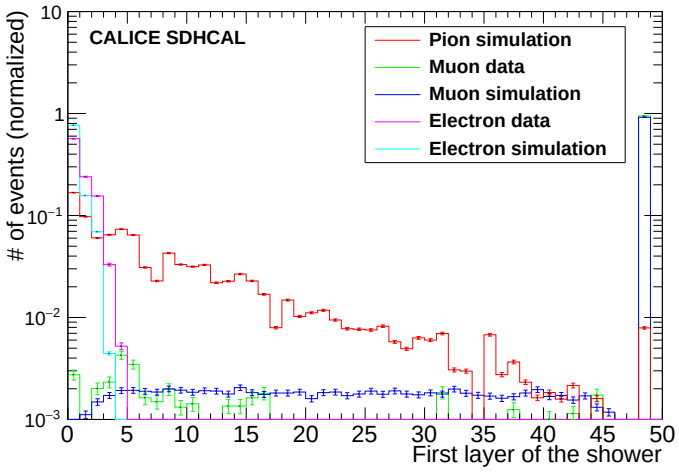

Figure 3. Distribution of the first layer of the shower (Begin). Layer 0 refers to the first layer of the prototype. Layer 48 is the virtual layer after the last layer and used to tag events not fulfilling first layer criteria. In the standard method described in ref. [7], events that start showering before the fifth layer are eliminated.

(even the radiative ones) from pions and electrons. It allows also a slight separation between pions and electrons.

- **Shower density (Density):** for each hit i, we count how many hits located in the 3×3 pads around it (including itself) to obtain N_i. The density is then defined as the average number of N_i following the formula: Density = $\sum_{i=1}^{N_{hit}} N_i/N_{hit}$, where N_{hit} is the total number of hits in the event. Figure 6 shows clearly that electromagnetic showers are more compact than the hadronic showers as expected.

- **Shower radius (Radius):** the RMS of each hit's distance from the event axis. To estimate the event axis, the average positions of the hits in each of the ten first fired layers of an event are used to fit a straight line. The straight line is then used as the event axis. Figure 7 shows the average radius of the three particle species in the SDHCAL. Discrepancy of the muon radius distribution between data and simulation is due to the difference of hit multiplicity which is slightly larger in data with respect to simulation.

- **Shower maximum position (Length):** this is the distance between the shower start and the layer where the maximum RMS of hit transverse coordimates with respect to the shower axis is detected. The distribution of this variable for different particle species is shown in figure 8.

Before using the variables listed above as input to the BDT method, we check that the variables distributions in the simulation are in agreement with data for the muon and electron beams which are quite pure. Figures 3–8 show that there is globally a good agreement for the six variables of the two species even though the agreement is not perfect in particular for electrons. The observed discrepancy is related to the difficulty to simulate precisely the saturation effect of electromagnetic showers in RPC detectors as explained in ref. [14].

2020 JINST 15 P10009

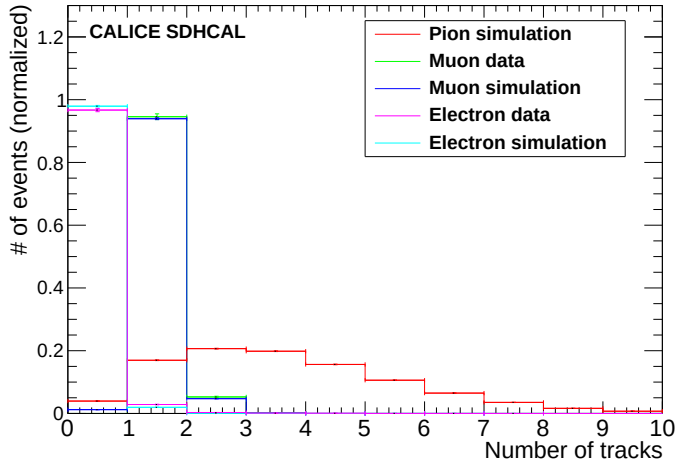

Figure 4. Distribution of number of the tracks in the shower (TrackMultiplicity).

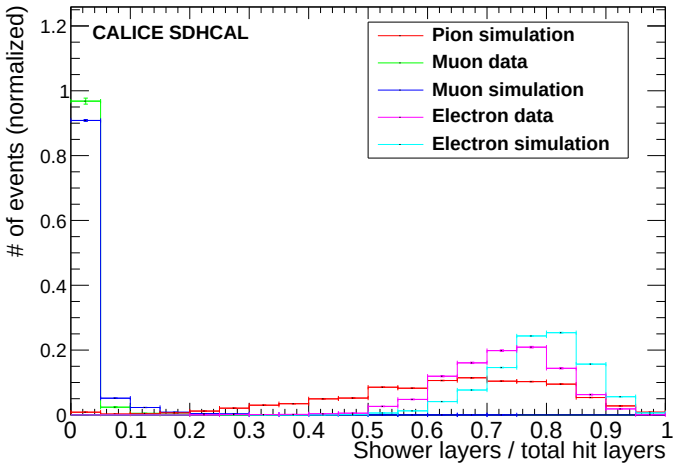

Figure 5. Distribution of ratio of the number of layers in which RMS of the hits' position in the x-y plane exceeds 5 cm over the total number of fired layers (NinteractingLayers/NLayers).

3.2 The two approaches to build the BDT-based classifier

In order to take into account the difference observed in some variable distributions between data and simulation, and to cross-check the particle identification using the BDT method, we adopt two different training strategies for the BDT-based classifier. The first approach, referred to as MC Training, uses simulation samples of pions, electrons and muons as training sets. The second, referred to as Data Training, uses simulation samples of pions but electron and muon samples taken from data as training sets.

2020 JINST 15 P10009

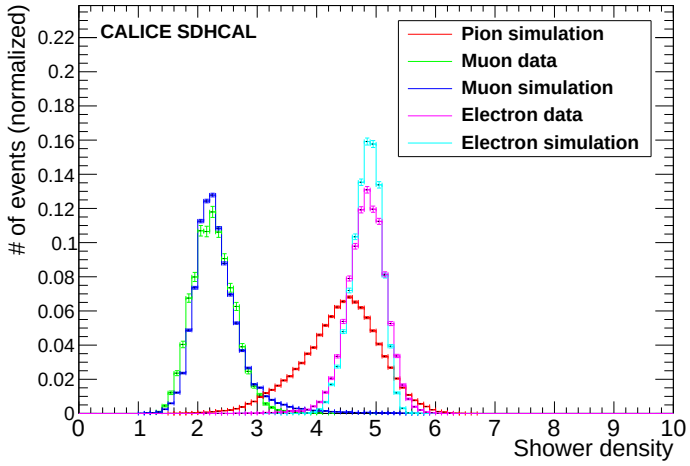

Figure 6. Distribution of the average number of neighbouring hits surrounding one hit (Density).

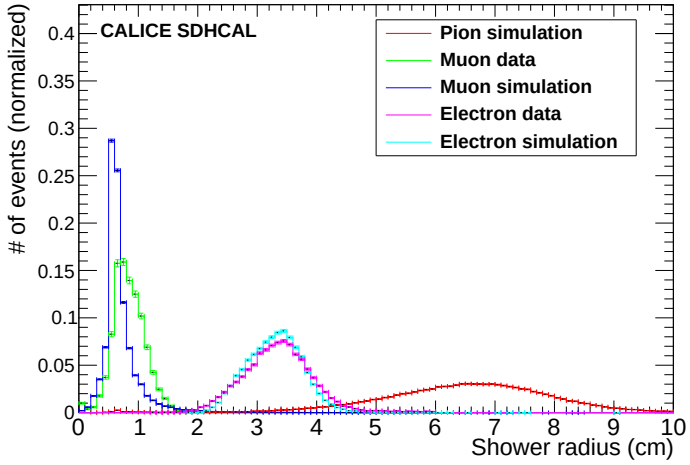

Figure 7. Distribution of the average radius of the shower (Radius).

3.2.1 MC training approach

The six variables of the simulated pion, muon and electron events described in section 3.1 are used for the training and testing of the classifier. Events are chosen in alternating turns for the training and test samples as they occur in the source trees until the desired numbers of training and test events are selected. The training and test samples contain the same number of events for each event class. Independent samples of signal events (pions) and of the different background contributions (electron and muons) are used. The ratio between signal and each background (electron or muon) events is 1 for training and test samples. After the training, the BDT provides the relative weight of each variable as a measure of distinguishing signal from background. Two BDT-based classifiers are proposed here. The first ($BDT_{\pi\mu}$) is used to discriminate pions against muons and the second ($BDT_{\pi e}$) to discriminate against electrons. Table 1 shows the variable

2020 JINST 15 P10009

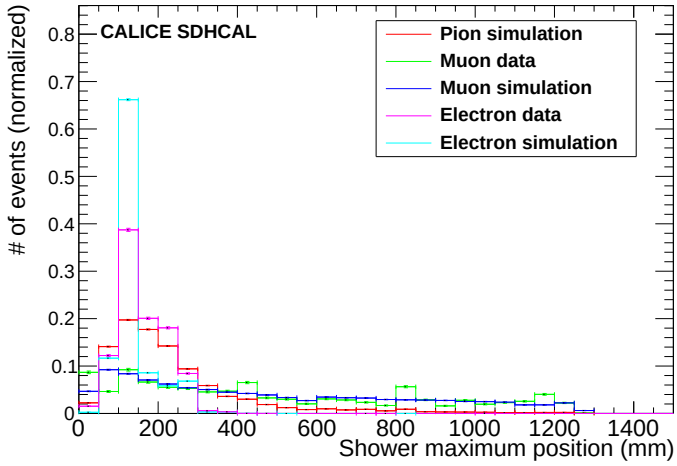

Figure 8. Distribution of the position of the layer with the maximum radius (Length).

Table 1. Variable ranking of separation power in the case of BDT$_{\pi\mu}$.

Rank : Variable	Variable relative weight
1 : Length	0.233
2 : Density	0.225
3 : NInteractinglayer/Nlayer	0.163
4 : Radius	0.160
5 : Begin	0.139
6 : TrackMultiplicity	0.080

ranking according to their separation power in the BDT$_{\pi\mu}$ while table 2 gives their separation power in the case of BDT$_{\pi e}$. The BDT algorithm using the variables and their respective weights is then applied to the test samples. The output of the BDT applied to each of the test sample events is a variable belonging to the interval [-1,1] with the positive value representing more signal-like events and the negative more background-like events.

Figure 9 (left) shows the output of the BDT for a test sample made of pions and muons while figure 9 (right) shows the output for a test sample made of pions and electrons. The values differ significantly for signal and background suggesting thus a large separation power of the BDT approach. This is confirmed by figure 10. The pion selection efficiency versus the muon (electron) rejection of the test sample is shown in figure 11 (left) and figure 11 (right), respectively. A pion selection efficiency exceeding 99.0% with a muon and electron rejection of the same level (> 99.0%) can be achieved. In order to check the validity of these two classifiers, we use the beam samples of pure muons and electrons. Figure 12 (left) shows the BDT output of BDT$_{\pi\mu}$ and figure 12 (right) shows the case of BDT$_{\pi e}$. Beam muon results show a good agreement with respect to the simulated

2020 JINST 15 P10009

Table 2. Variable ranking of separation power in the case of BDT$_{\pi e}$.

Rank : Variable	Variable relative weight
1 : Radius	0.204
2 : NInteractinglayer/Nlayer	0.203
3 : Density	0.194
4 : Length	0.151
5 : Begin	0.145
6 : TrackMultiplicity	0.101

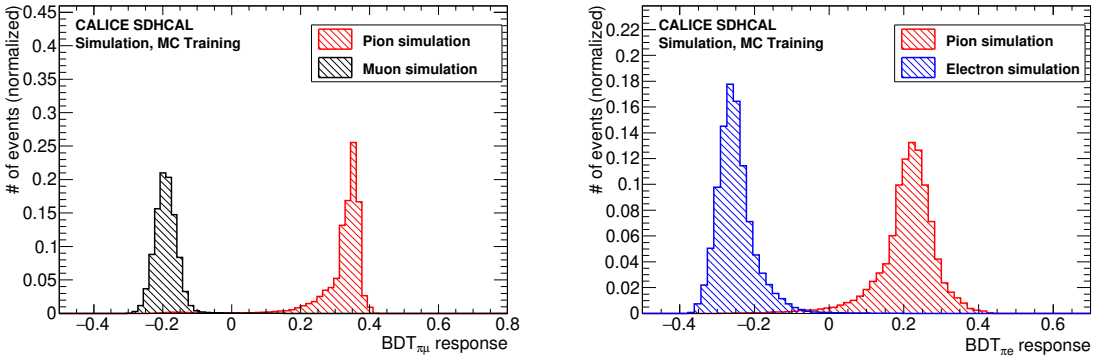

Figure 9. The BDT output of the BDT$_{\pi\mu}$ (left) and BDT$_{\pi e}$ (right) built with simulation samples.

events. A shift of the beam electron shape is observed with respect to the one obtained from the simulated events. This difference is most probably due to the fact that the distribution of some variables in data and in the simulation are not identical. Next, as a first step in purifying the collected hadronic data events we apply the pion-muon classifier. Figure 12 (left) shows the BDT$_{\pi\mu}$ response applied to the collected hadron events in the SDHCAL. We can clearly see that there are two maxima. One maximum in the muon range corresponds to the muon contamination of pion data and another one in the pion range. Hence, to ensure the rejection of the muons in the sample, the BDT variable is required to be > 0.1. The second step is to apply the BDT$_{\pi e}$ to the remaining of the pion sample. Figure 12 (right) shows the BDT$_{\pi e}$ output. In order to eliminate the maximum of the electrons contamination and get almost a pure (> 99.5%) pion sample with limited loss of pion events, we apply to the pion samples a BDT$_{\pi e}$ cut of 0.05.

3.2.2 Data training approach

We use the same variables of the MC Training approach on the data samples of muons (11k events) and electrons (30k events) but still on the simulated pion samples to build two classifiers. Then we apply the same procedure as the MC Training approach. Table 3 and 4 show the corresponding variables ranking for BDT$_{\pi\mu}$ and BDT$_{\pi e}$ according to their separation power importance. The difference of variables weights of these two tables with respect to those obtained with MC training

2020 JINST 15 P10009

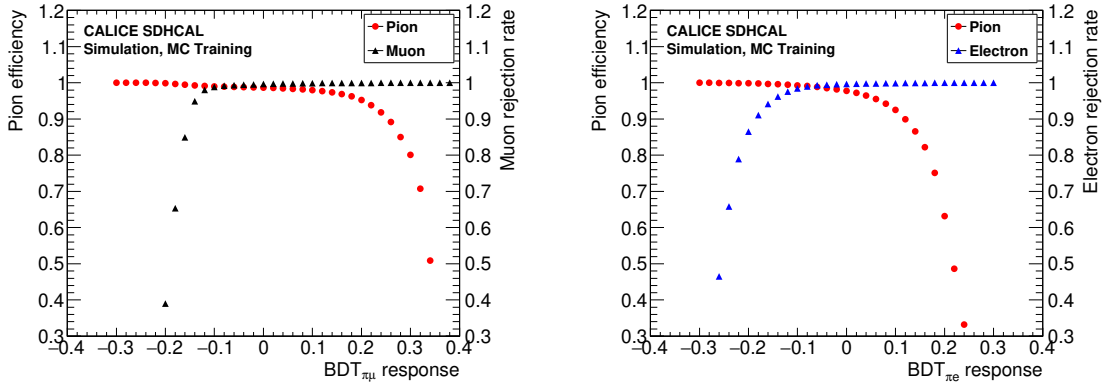

Figure 10. Pion efficiency and muon rejection rate (left) and pion efficiency and electron rejection rate (right) as a function of the BDT output.

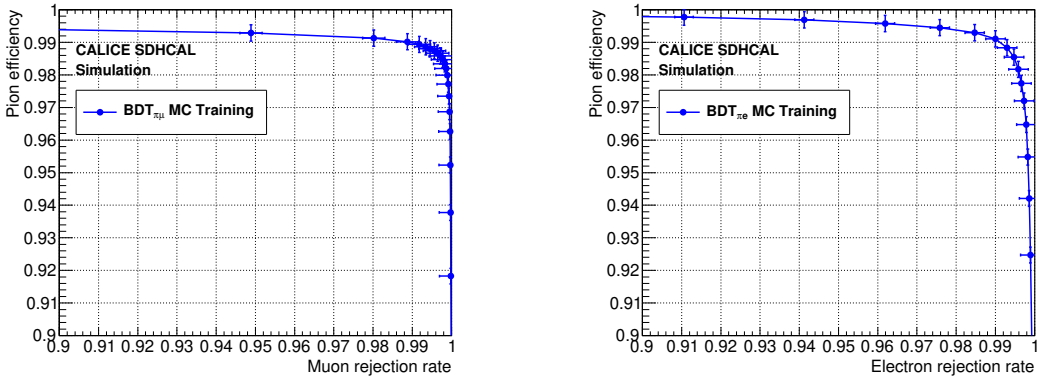

Figure 11. Pion efficiency versus muon rejection rate(left) and pion efficiency versus electron rejection rate (right).

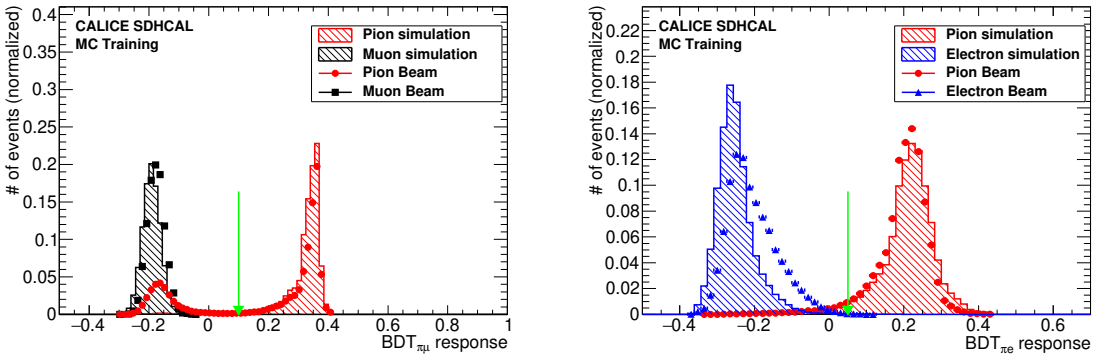

Figure 12. The BDT output after using the BDT$_{\pi\mu}$ on the data pion sample (left) and the BDT output after using the BDT$_{\pi e}$ on the same data pion sample after classified by BDT$_{\pi\mu}$ (right). A green arrow is shown on both to indicate the BDT cut applied to clean the pion samples.

2020 JINST 15 P10009

Table 3. Variable ranking of separation importance in the case of $BDT_{\pi\mu}$.

Rank : Variable	Variable relative weight
1 : Length	0.300
2 : Radius	0.230
3 : Density	0.227
4 : Begin	0.103
5 : NInteractinglayer/Nlayer	0.080
6 : TrackMultiplicity	0.060

Table 4. Variable ranking of separation importance in the case of $BDT_{\pi e}$.

Rank : Variable	Variable relative weight
1 : Radius	0.195
2 : NInteractinglayer/Nlayer	0.191
3 : Density	0.189
4 : Length	0.151
5 : Begin	0.141
6 : TrackMultiplicity	0.131

approach is explained by the slight difference of some variables distributions between data and simulation. Indeed, when dropping, in the BDT method, the variables for which the discrepancy between data and simulation is present, namely the "Begin" and "NinteractingLayers/NLayers" variables, similar weights are obtained for the remaining variables in the two approaches. Figure 13 left (right) gives the results of pion efficiency and muon (electron) rejection rate. The left (right) plot of figure 14 shows the BDT output of the $BDT_{\pi\mu}$ ($BDT_{\pi e}$). Clearly these two classifiers have very good separation power. We apply these classifiers to the raw pion beam samples. The results can be seen in figure 15. We apply a BDT cut value of 0.2 in the pion-muon separation stage and then a BDT cut value of 0.05 in the pion-electron separation stage.

4 Results

The distributions of input variables for the data and simulation events of pion, muon and electron are shown in figure 16. Only the pion data sample distributions are obtained after applying the data-based BDT classifiers. A good agreement between the data and simulation events for pions is observed. This confirms the power of the BDT method. The muon rejection rate obtained with the MC (Data) Training approach is found to be 99.9% (99.0%) and that of the electron is 99.8% (99.4%) respectively. The difference between data and simulation using the two BDT training

2020 JINST 15 P10009

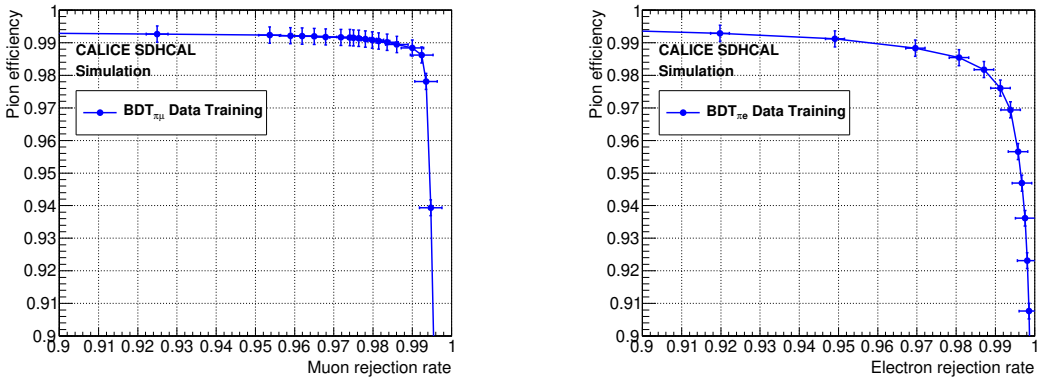

Figure 13. Pion efficiency versus muon rejection rate (left) and pion efficiency versus electron rejection rate (right).

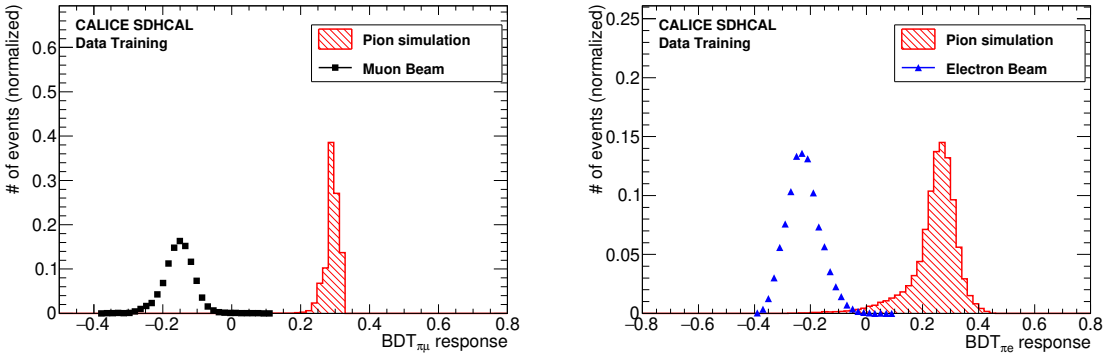

Figure 14. BDT output of the BDT$_{\pi\mu}$ built with pure beam muons and simulated pion samples (left) and of the BDT$_{\pi e}$ built with pure beam electrons and simulated pion samples (right).

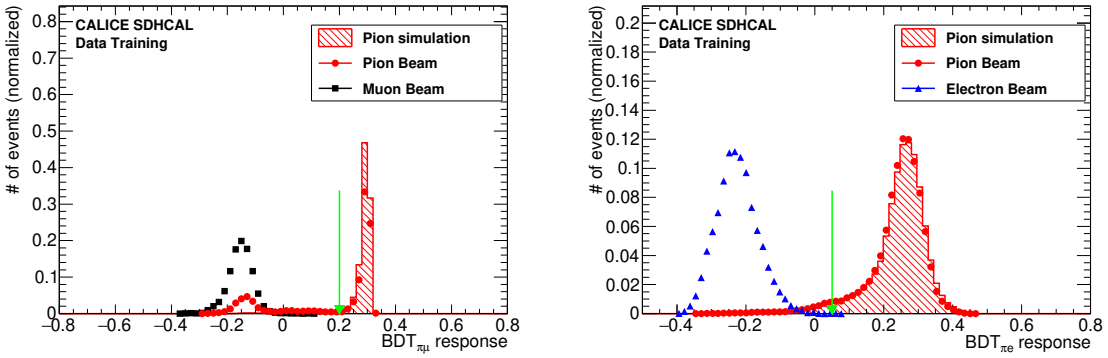

Figure 15. The BDT output after using the BDT$_{\pi\mu}$ on the data pion sample (left) and the BDT output after using the BDT$_{\pi e}$ on the same pion sample after classified by BDT$_{\pi\mu}$ (right). A green arrow is shown on both to indicate the BDT cut applied to clean the pion samples.

2020 JINST 15 P10009

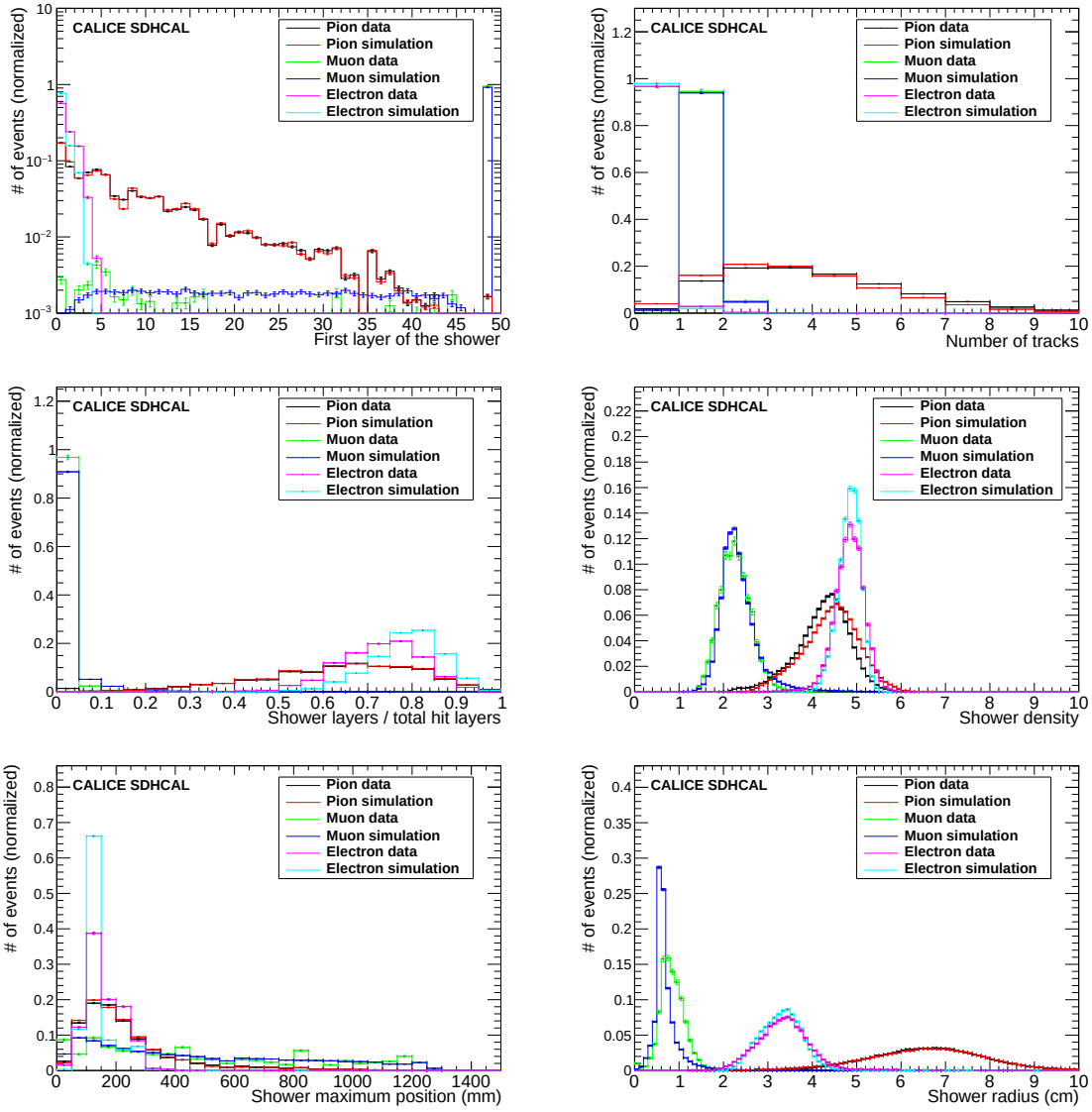

Figure 16. Distributions of six input variables of electron, muon and pion samples. Pion distributions are obtained from pion data samples after applying the data-based training BDT selections.

approches for different BDT cuts in the range of [0.05,0.25] in the case of the pion-muon separation shows a value smaller than 0.2% in both the pion efficiency and the muon rejection rate. In the case of the pion-electron separation, different BDT cuts in the range of [0.05, 0.15] result in a difference of less than 1.2% in the pion efficiency and less than 0.2 % in the electron rejection rate.

The rejection of muons and electrons presented in the pion data sample using the BDT allows us to have a rather pure pion sample as explained in the previous section. Figure 17 shows the results of comparison in event selection between the standard method and the BDT-based method using the simulation samples. For both simulation and beam data, the BDT method leads to a larger pure sample of hadrons comparing to the standard method [7] in particular at low energy as shown in figure 18 for the comparison of the selected events as a function of the total number of hits for the 10 GeV pion

2020 JINST 15 P10009

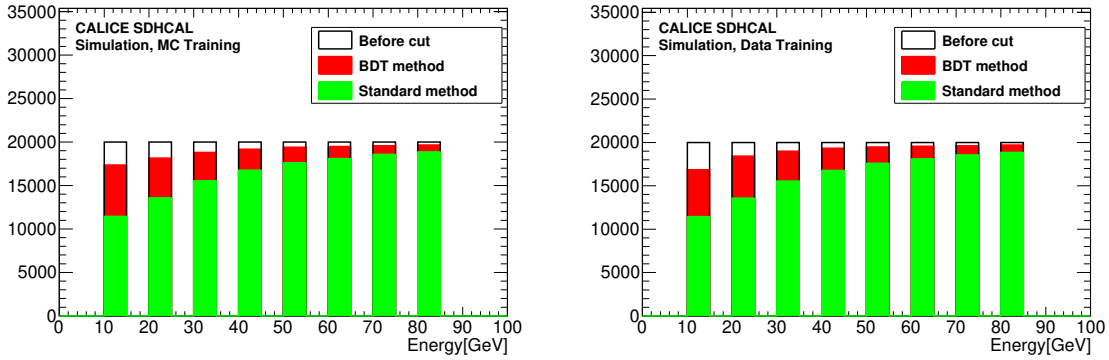

Figure 17. The number of simulated events of different energy points from 10 GeV to 80 GeV before (white) and after applying the standard method ref. [7] (green) or BDT method (red). The left plot shows the results from BDT method with MC Training approach while the right one shows the results with Data Training approach.

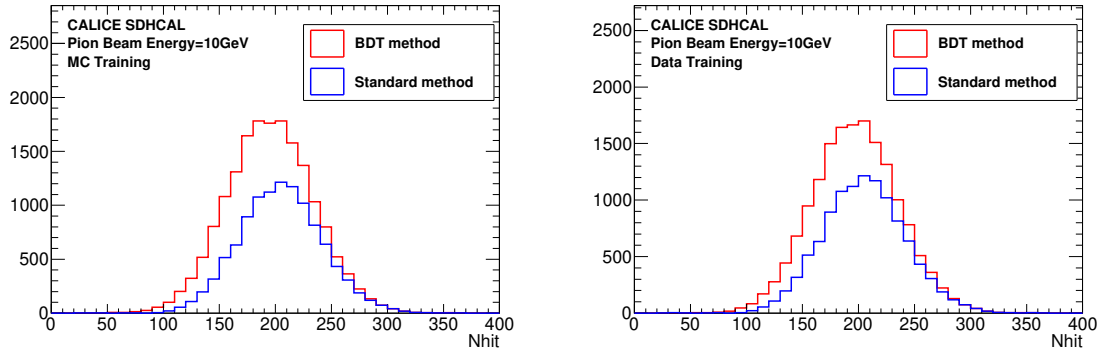

Figure 18. Distribution of the total number of hits for the 10 GeV pion beam data selected by the standard method (blue) and the BDT method (red). The left plot shows the results from BDT method with MC Training approach while the right one shows the results with Data Training approach.

beam data. We also do not observe any significant deviation of energy resolution when applying the standard energy reconstruction described in ref. [7] on the pion events selected by the BDT method.

5 Conclusion

A new particle identification method using BDT-based MVA technique is applied to purify the pion events collected at the SPS H2 beamline in 2015 by the CALICE SDHCAL prototype. The new method uses the topological shape of events associated to muons, electrons and pions in the CALICE SDHCAL to reject the two first species. A rejection rate of muons (electrons) exceeding 99.0% (99.4%) respectively with a pion selection efficiency of about 95.0% is obtained. A significant statistical gain is obtained with respect to the standard method used in the work presented in ref. [7]. This statistical gain is particularly significant at energies up to 40 GeV and can be explained by the fact that the showers that start in the first layers are not all rejected. This gain shows the better efficiency and separation power of the multivariate approach over the cut-based approach of the

2020 JINST 15 P10009

standard method. The BDT-based particle identification in CALICE SDHCAL is a robust and a reliable method as confirmed by the results of two different training approaches. Finally, a study of the linearity and the resolution of the reconstructed energy of the hadronic showers selected by the BDT-based method in the SDHCAL is in preparation and will be the object of a future paper.

Acknowledgments

This study was supported by National Key Programme for S&T Research and Development (Grant NO. 2016YFA0400400).

References

[1] G. Baulieu et al., *Construction and commissioning of a technological prototype of a high-granularity semi-digital hadronic calorimeter*, 2015 *JINST* **10** P10039 [arXiv:1506.05316].

[2] Linear Collider ILD Concept Group collaboration, *The International Large Detector: letter of intent*, FERMILAB-LOI-2010-01, (2010) [FERMILAB-PUB-09-682-E] [DESY-2009-87] [KEK-REPORT-2009-6] [arXiv:1006.3396].

[3] M.A. Thomson, *Particle flow calorimetry and the PandoraPFA algorithm*, *Nucl. Instrum. Meth. A* **611** (2009) 25 [arXiv:0907.3577].

[4] V.L. Morgunov, *Calorimetry design with energy-flow concept (imaging detector for high-energy physics)*, in *Proc. of Int. Conf. on Calorimetry (Calor*02*)*, World Scientific, Singapore (2003).

[5] J.-C. Brient and H. Videau, *The calorimetry at the future e^+e^- linear collider*, in *Proc. of APS/DPF/DBP summer study on the future of particle physics*, Snowmass, CO, U.S.A. (2002) [*eConf* **C 010630** (2001) E3047] [hep-ex/0202004].

[6] F. Dulucq, C. de La Taille, G. Martin-Chassard and N. Seguin-Moreau, *HARDROC: readout chip for CALICE/EUDET digital hadronic calorimeter*, *IEEE Nucl. Sci. Symp. Med. Imag. Conf. Rec.* (2010) 1678.

[7] CALICE collaboration, *First results of the CALICE SDHCAL technological prototype*, 2016 *JINST* **11** P04001 [arXiv:1602.02276].

[8] B.P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu and G. McGregor, *Boosted decision trees, an alternative to artificial neural networks*, *Nucl. Instrum. Meth. A* **543** (2005) 577 [physics/0408124].

[9] H.-J. Yang, B.P. Roe and J. Zhu, *Studies of boosted decision trees for MiniBooNE particle identification*, *Nucl. Instrum. Meth. A* **555** (2005) 370 [physics/0508045].

[10] A. Hocker et al., *TMVA — toolkit for multivariate data analysis*, physics/0703039.

[11] GEANT4 collaboration, *GEANT*4 *— a simulation toolkit*, *Nucl. Instrum. Meth. A* **506** (2003) 250.

[12] V.V. Uzhinsky, *The Fritiof (FTF) model in GEANT*4, in *Proceedings, International Conference on Calorimetry for the High Energy Frontier (CHEF* 2013*)*, (2013), pg. 260.

[13] G. Folger, V.N. Ivanchenko and J.P. Wellisch, *The binary cascade*, *Eur. Phys. J. A* **21** (2004) 407.

[14] CALICE collaboration, *Resistive plate chamber digitization in a hadronic shower environment*, 2016 *JINST* **11** P06014 [arXiv:1604.04550].

[15] CALICE collaboration, *Tracking within hadronic showers in the CALICE SDHCAL prototype using a Hough transform technique*, 2017 *JINST* **12** P05009 [arXiv:1702.08082].

2020 JINST 15 P10009